Visual Classification via Description from Large Language Models

Year: 2022 | Citations: 359 | Authors: Sachit Menon, Carl Vondrick

---

Abstract

Vision-language models (VLMs) such as CLIP have shown promising performance on a variety of recognition tasks using the standard zero-shot classification procedure -- computing similarity between the query image and the embedded words for each category. By only using the category name, they neglect to make use of the rich context of additional information that language affords. The procedure gives no intermediate understanding of why a category is chosen, and furthermore provides no mechanism for adjusting the criteria used towards this decision. We present an alternative framework for classification with VLMs, which we call classification by description. We ask VLMs to check for descriptive features rather than broad categories: to find a tiger, look for its stripes; its claws; and more. By basing decisions on these descriptors, we can provide additional cues that encourage using the features we want to be used. In the process, we can get a clear idea of what features the model uses to construct its decision; it gains some level of inherent explainability. We query large language models (e.g., GPT-3) for these descriptors to obtain them in a scalable way. Extensive experiments show our framework has numerous advantages past interpretability. We show improvements in accuracy on ImageNet across distribution shifts; demonstrate the ability to adapt VLMs to recognize concepts unseen during training; and illustrate how descriptors can be edited to effectively mitigate bias compared to the baseline.